S^{2}Net: Skeleton-aware SlowFast Network for Efficient Sign Language Recognition

Yifan Yang[1], Yuecong Min[2,3], and Xilin Chen[2,3]

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
yyf_355@hust.edu.cn, {minyuecong, xlchen}@ict.ac.cn

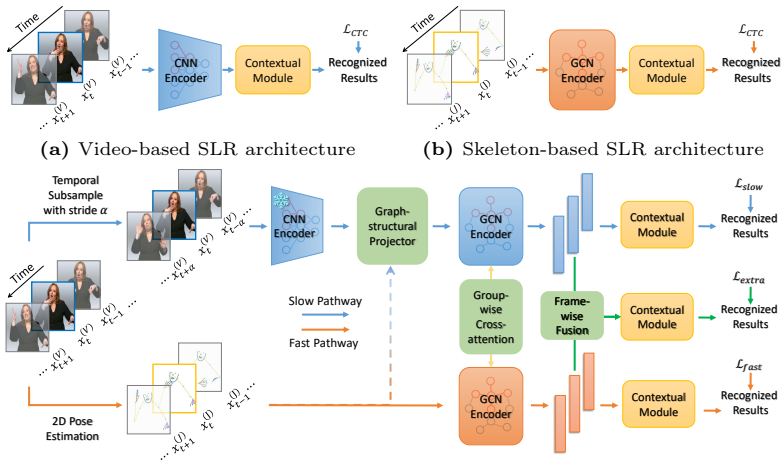
Abstract. Continuous Sign Language Recognition (CSLR) aims to interpret meaning from signers' postures and movements. Joint-wise correspondences between estimated skeleton data and sign videos provide complementary insights into appearance and motion. In this paper, we propose a Skeleton-aware SlowFast Network(S^{2}Net) to effectively capture the appearance and motion information in sign videos. S^{2}Net leverages skeleton data in the fast pathway and video data in the slow pathway, progressively integrating both streams of information. Initially, we project both skeleton and video data into a unified graph-structured space and employ a consistent GCN-based architecture for both pathways, then we propose a group-wise cross-attention module to fuse intermediate features between different pathways. Finally, a frame-wise fusion pathway is adopted to integrate the semantic information at the sequence level. Experimental results on three public datasets demonstrate the effectiveness and efficiency of the proposed method.

Keywords: Sign Language Recognition · Nonverbal Communication · Graph Neural Networks · Human Computer Interaction

1 Introduction

As a visual language, sign language conveys information through dynamic movements like gestures and facial expressions, as well as static handshapes and positions, collectively forming a rich and nuanced mode of communication. Continuous Sign Language Recognition (CSLR) aims to automatically identify and interpret signs from continuous video streams, which plays a critical role in facilitating communication for the hearing impaired and bridging language barriers. By enabling real-time transcription of sign language into text or speech, CSLR not only enhances inclusion but also empowers individuals to participate more fully in educational, professional, and social environments.

Sign language relies heavily on hand, face, and body movements, necessitating the extraction of detailed information. Recent strides in deep learning have significantly enhanced the accuracy of pose estimation systems [4, 6, 20],



(a) Video-based SLR architecture **(b)** Skeleton-based SLR architecture

(c) The proposed Skeleton-aware SlowFast Network (S^{2}Net)

Fig. 1: Illustration of (a) video-based and (b) skeleton-based CSLR framework, and (c) the proposed S^{2}Net. S^{2}Net leverages video and skeleton data with different frame rates as inputs and adopts consistent GCN-based architecture. The video data is temporally down-sampled and projected to a graph-structured space, which is further integrated with skeleton data through group-wise cross-attention and frame-wise fusion.

prompting numerous studies to explore how pose information can guide the learning process of CSLR models. Some works [56, 57] integrate an additional pose estimation branch or directly utilize pose data to enrich visual representations. Other methods [9, 36] explore the incorporation of various visual inputs, such as skeleton data and optical flow. However, methods that represent skeleton data as 3D heatmaps or adopt late fusion strategies often compromise the computational efficiency of skeleton data.

Skeleton-based action recognition [43] has gained considerable attention from researchers for its efficiency, appearance invariance, and privacy preservation. Recent CSLR studies [9, 27] have shown that skeleton-based methods can achieve competitive results with video-based methods while maintaining better efficiency. However, relying solely on skeleton data can lead to irreversible information loss, and skeleton-based CSLR methods often struggle with accurately recognizing signs due to inaccurate estimation, especially during fast hand movements or instances of occlusion. In this paper, we propose a Skeleton-aware SlowFast Network (S^{2}Net) to efficiently integrate motion information from skeleton data and appearance information from video data.

Instead of representing skeleton data as 3D heatmaps [9, 15], we project the feature maps extracted from video data into a graph-structured space, preserving detailed visual information while enhancing efficiency. As shown in Fig. 1c, S^{2}Net leverages skeleton data for the fast pathway and video data for the slow pathway, progressively integrating both types of information. Specifically, we initially project both skeleton and video data into a unified graph-structured

space and employ a consistent GCN-based architecture for both pathways, then we propose a group-wise cross-attention module to fuse intermediate features of different pathways. A frame-wise fusion pathway is adopted to fuse the semantic information at the sequence level. To better explore the transferability of the learned features, we evaluate the effects of distilling the ensemble prediction from the trained S^2Net to a skeleton-only model.

In conclusion, this paper leverages the complementary characteristics of video and skeleton data in CSLR and proposes S^2Net to integrate both types of information progressively. Experimental results on three public CSLR datasets demonstrate the effectiveness of the proposed method, achieving significant performance gains with reduced training costs.

The main contributions are summarized as follows:

- ◇ Projecting the feature maps extracted from video data into a graph-structured space to preserve detailed visual information while enhancing efficiency.
- ◇ Proposing a Skeleton-aware SlowFast Fusion method for the progressive fusion of skeleton and video information.
- ◇ Evaluating the transferability of ensemble predictions from the trained S^2Net to a skeleton-only model.

2 Related Work

2.1 Sign Language Recognition

Sign language recognition can be roughly categorized into isolated sign language recognition (ISLR) [24, 28, 33, 51, 52] and continuous sign language recognition (CSLR) [11, 13, 27, 30, 34, 38], which aim to identify signs occurring in segmented or unsegmented videos and have seen significant advancements. Earlier efforts for CSLR focus on utilizing weakly annotated data, employing hybrid learning methods [30–32] and pure deep learning techniques [10, 11, 13, 34, 38]. DNF [13] adopts an iterative training mechanism to capture discriminative visual features. Approaches like VAC [34] further achieve competitive performance in an end-to-end fashion, enhancing the efficiency of the training process. A typical sign language framework [13, 34, 35] comprises a visual feature extractor, a contextual module, and an alignment module. Considering the training efficiency, recent works [26, 50, 53] often employ frame-wise visual extractor (e.g., ResNet [23], and ViT [14]), and some works [2, 9, 47, 58] utilize pre-trained spatio-temporal visual extractor like I3D [7] and S3D [48]. Compared to image-based methods, skeleton-based methods [27, 36, 56, 57] are more efficient during training but require an additional pose estimation stage. CoSign [27] focuses on the utilization of co-occurrence signals and employs group-specific Graph Convolutional Networks (GCN) to capture visual features in skeleton sequences. Due to the limited temporal receptive field of the visual feature extractor, most works [9, 21, 22, 27, 53] integrate RNNs or Transformers to capture contextual information. Some works [21, 22, 53] further exploit the relationship between the

visual extractor and the contextual module. Connectionist Temporal Classification (CTC) [19] is widely used to provide alignment supervision based on sentence-wise annotations due to its efficiency.

In addition to exploiting visual information alone, recent works attempt to enhance visual features by incorporating additional information. As a typical visual language, some works [2, 47, 53, 54, 58] try to leverage the textual information to guide the learning of visual features. Albanie *et al.* [2] build a large-scale British sign language dataset named BSL-1k with co-articulated mouthing cues. NLA-SLR [58] leverages semantic information from glosses to guide the learning of visual representation. To better exploit body-relevant information in sign videos, STMC [56] adopts an extra pose estimation branch to guide the learning of visual features. C^2SLR [57] leverages pose heatmaps to guide the visual module to focus on informative regions. SlowFastSign [1] utilizes the two-pathway SlowFast network to capture spatial and dynamic features at different temporal resolutions with a bi-directional feature fusion. TwoStream-SLR [9] adopts a two-stream framework to gradually fuse visual features from RGB videos and keypoint heatmaps. Different from TwoStream-SLR, we directly capture the dynamic information from skeleton sequences, instead of heatmaps, and fuse the visual and dynamic information in graph-structured space with GCN-based methods efficiently.

2.2 Skeleton-based Action Recognition

With the development of pose estimation technology [4, 6, 20], skeleton-based action recognition has garnered significant attention in recent years due to its robustness to illumination changes and scene variation with high efficiency [43]. Various approaches have been proposed to address the challenges posed by this task. Amir*et al.* [41] collects a large-scale RGB-D dataset for 3D human activity analysis and proposes a part-aware LSTM to keep the context of each body part independent. ST-GCN [49] leverages the graph structure and designs spatial-temporal graph convolution operation to learn human representations from 2D or 3D coordinates. PoseConv3D [15] revisits the GCN-based methods and proposes to use 3D heatmap volume instead of graph sequence as input to improve the robustness, interoperability, and scalability. In contrast, we propose to project visual representation into the graph-structured space, which can reduce redundant information and improve recognition efficiency.

2.3 Video Representation Learning

Video representation learning [40] has garnered significant attention due to its crucial role in various computer vision tasks such as action recognition, video captioning, and anomaly detection. Many efforts have been made to extend the success of 2D CNNs into the temporal domain, including the use of inflated pre-trained weights [7], the space-time decomposition [16, 39, 48], attention designs [3, 44], multi-stream fusion [17, 42], and self-supervised pre-training [18, 45, 46]. Different from the image-based task, video representation learning needs to

capture both appearance and motion information. Two-stream [42] achieves this by extracting appearance information from individual frames and motion information from the optical flow between frames, effectively encoding both cues by integrating these streams. Similarly, SlowFast [1] adopts a two-pathway design: a slow pathway with a low framerate for detailed spatial details and a fast pathway with a high framerate for dynamic motion cues. Video representation learning requires processing complex spatio-temporal information, demanding substantial computational resources for both training and inference. X3D [16] progressively expands a 2D architecture into a spatio-temporal one considering factors like temporal duration, framerate, and spatial resolution. Feichtenhofer *et al.* [18] extend MAE to video and observe higher masking ratio improve both efficiency and accuracy. Inspired by these works, we project the visual representation into the graph-structured space and fuse the visual and skeleton information with different frame rates to better balance efficiency and accuracy.

3 Method

This section first outlines the basic CSLR pipeline in Sect. 3.1, followed by the projection of visual features into a graph-structured representation in Sect. 3.2. Next, we propose a Skeleton-aware SlowFast Fusion method in Sect. 3.3. Finally, we design a multi-head distillation method in Sect. 3.4.

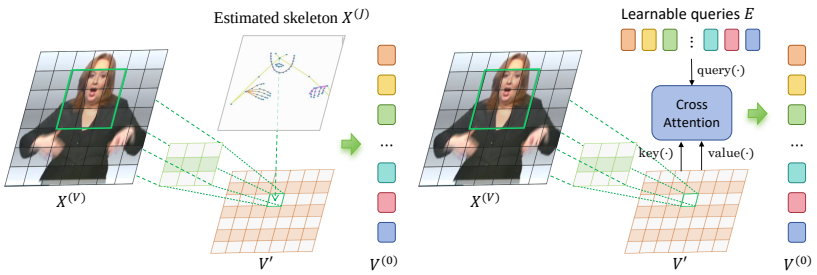
3.1 Preliminaries

Video-based CSLR aims to identify the sign sequence $\boldsymbol{l} = (l_1, \cdots, l_N)$ from a sign video $\boldsymbol{X}^{(V)} = (\boldsymbol{x}_1^{(V)}, \cdots, \boldsymbol{x}_T^{(V)}) \in \mathbb{R}^{T\times H\times W\times 3}$, where N is the number of signs, and T, H and W are the video's length, height and width, respectively, and the superscript denotes the video input. Similar to previous work [34], we first extract frame-wise features $\boldsymbol{V} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_T)$ with a pre-trained ResNet18, then aggregateing contextual information via 1D-CNN and BiLSTM layers and make prediction $\boldsymbol{y}^{(V)} = (\boldsymbol{y}_1^{(V)}, \cdots, \boldsymbol{y}_{T'}^{(V)})$, where $\boldsymbol{y}^{(V)}$ may have a different length due to temporal subsampling. CTC is adopted to supervise the monotonous alignments between predictions $\boldsymbol{y}$ and gloss sequences $\boldsymbol{l}$:

$$\mathcal{L}_{\text{CTC}} = -\log p(\boldsymbol{l}|\boldsymbol{y}^{(V)}) = -\log_{\pi\in\mathcal{B}^{(-1)}(\boldsymbol{l})} p(\pi|\boldsymbol{y}^{(V)}), \tag{1}$$

where π is a frame-wise gloss sequence that correspondent to $\boldsymbol{y}$, and $\mathcal{B}$ builds a many-to-one mapping from π to the correspondence gloss sequence $\boldsymbol{l}$.

With the rapid development of pose estimation techniques [4, 6, 20], real-time estimation of skeletons can provide sufficient information for CSLR. We employ the Group-specific GCN module proposed in previous work [27] to extract frame-wise features $\boldsymbol{J} = (\boldsymbol{j}_1, \cdots, \boldsymbol{j}_T)$ from the estimated skeleton sequence $\boldsymbol{X}^{(J)} = (\boldsymbol{x}_1^{(J)}, \cdots, \boldsymbol{x}_T^{(J)}) \in \mathbb{R}^{T\times K\times 3}$, where K denotes the number of joints per skeleton and $\boldsymbol{X}^{(J)}$ consists of 2D joint coordinates and confidence scores. As shown in Fig. 1a and 1b, we adopt a similar design for the rest of the



(a) Index-based projector **(b)** Query-based projector

Fig. 2: Illustration of the proposed Graph-structured Projectors.

skeleton-based framework as the video-based framework. While the skeleton-based approach is more efficient, it suffers from information loss during pose estimation, whereas the video-based method retains more visual details but is less efficient. We leverage the strengths of both frameworks to strike a balance between efficiency and accuracy.

3.2 Graph-structured Representation

Skeleton data offers efficient and interpretable posture and motion information through coordinate vectors. Unlike previous works [9,15] that transform skeleton data into 3D heatmaps, we project video data into a graph-structured representation for a unified framework. As shown in Fig. 1c, we first use an off-the-shelf pose estimator [12] to estimate 2D joints $\boldsymbol{X}^{(J)}$, which includes hand, mouth, face, and body parts to model both manual and non-manual signals in sign language. Similar to previous work [27], the estimated joints are grouped into five parts based on human anatomy and projected to the graph-structured space:

$$\boldsymbol{J}^{(0)} = f(\boldsymbol{X}^{(J)}), \tag{2}$$

where $f(\cdot)$ is a linear projection layer with activation, the superscript of $\boldsymbol{J}$ indicates the number of feature extraction modules passed through, and 0 denotes the input for the feature extractor.

Although estimated skeleton data preserve most structural information, inaccurate joints may introduce noises and irreversible information loss. Therefore, we leverage the original video data to provide complementary visual details. For efficiency and interoperability, we extract mid-level visual features $\boldsymbol{V'} \in \mathbb{R}^{T\times C\times \eta H\times \eta W}$ from the shallow layers of a pre-trained model (*e.g.*, a ResNet18), where η is the spatial downsampling ratio to control the granularity. We design two kinds of projectors to obtain graph-structured representation $\boldsymbol{V^{(0)}}$: a query-based projector uses learnable joint-wise queries, while an index-based projector relies on the estimated skeleton data.

Index-based Projector. A logical approach is to directly index the extracted feature map $\boldsymbol{V'}$ based on the estimated skeleton $\boldsymbol{J}$, as illustrated in

Fig. 2a, the spatial context within $\boldsymbol{V'}$ helps mitigate inaccuracies in the skeleton estimation. This process can be formulated as:

$$\boldsymbol{V^{(0)}} = g(\boldsymbol{V'}; \boldsymbol{X}^{(J)}), \tag{3}$$

where $g(\cdot;\cdot)$ first finds the closest element in $\boldsymbol{V'}$ for each joint in $X^{(J)}$, and then project them into the same dimension with $\boldsymbol{J}^{(0)}$.

Query-based Projector. We also explore building graph-structured representation without the utilization of the estimated skeleton. As shown in Fig. 2b, we utilize K learnable embeddings $\boldsymbol{E} \in \mathbb{R}^{K\times C}$ to aggregate relevant information from the extracted feature map $\boldsymbol{V'}$ and assume each learnable query has the ability to identify the corresponding region through the cross-attention between it and the extracted feature maps. Subsequently, we organize the captured features into a graph sequence:

$$\boldsymbol{V}^{(0)} = h(\boldsymbol{E}, \boldsymbol{V'}), \tag{4}$$

where $h(\cdot,\cdot)$ first aggregates relevant elements from $\boldsymbol{V'}$ through cross attention, and then projects them into the same dimension with $\boldsymbol{J}^{(0)}$.

With the proposed projectors above, we can project both skeleton and video data ($\boldsymbol{X}^{(J)}$ and $\boldsymbol{X}^{(V)}$) into graph-structured representations ($\boldsymbol{J}^{(0)}$ and $\boldsymbol{V}^{(0)}$), enhancing the interoperability between branches. We can use the same architecture for both types of data, and design fine-grained interaction between them.

3.3 Skeleton-aware SlowFast Fusion

Sign language conveys information through both static information (*e.g.*, body posture and facial expressions) and dynamic information (*e.g.*, hand gestures and mouth shape). Therefore, we propose a Skeleton-aware SlowFast Fusion approach to gradually integrate dynamic information from skeleton data with static information from video data.

Unlike the original SlowFast [17], which uses video frames for both pathways, we utilize skeleton data for the fast pathway and video data for the slow pathway. Specifically, the fast pathway learns dynamic information from projected skeleton feature $\boldsymbol{J^{(0)}}$ with T frames, while the slow pathway learns fine-grained appearance information from projected feature $\boldsymbol{V}^{(0)}$ with a temporal stride of α, controlling the framerate ratio between two pathways. We further design a Skeleton-aware SlowFast Fusion approach to progressively integrate the information from two pathways.

Basic Architecture Design. Similar to previous work [27, 49], we group the estimated joints into five parts based on human anatomy and employ the same group-specific GCN. The graph aggregation process for the fast pathway on joints of group g can be formulated as:

$$\begin{aligned} \tilde{\boldsymbol{J}}_{g}^{(i)} &= \sum_{k} \Lambda_{g,k}^{-\frac{1}{2}} A_{g,k} \Lambda_{g,k}^{-\frac{1}{2}} \boldsymbol{J}^{(i)} W_{g,k}^{s} \\ \boldsymbol{J}_{g,t}^{(i+1)} &= \sum_{\tau=0}^{\Gamma} \tilde{\boldsymbol{J}}_{g,t-\lfloor \Gamma/2 \rfloor+\tau}^{(i)} W_{\tau}^{t} + \boldsymbol{J}_{g,t}^{(i)} W^{i} \end{aligned} \tag{5}$$

where $\boldsymbol{V}^{(i)}$ denotes the input features for the ith GCN layer, and $\boldsymbol{V}^{(i+1)}$ denotes the output features. $A_{g,k}$ and $\Lambda_{g,k}$ are the adjacency matrix and degree matrix of group g for the participation k. Γ denotes the temporal receptive field. The weight matrices W^s, W^t and W^i correspond to spatial convolution, temporal convolution and identify projection, respectively. After obtaining the group-specific features, we employ a contextual module consisting of a temporal convolution and BiLSTM layer to extract contextual information. The basic supervision is:

$$\mathcal{L}_{basic} = \mathcal{L}_{CTC}(\boldsymbol{y}_a^{(J)}) + \mathcal{L}_{CTC}(\boldsymbol{y}_p^{(J)}) + \lambda_{CR}\mathcal{L}_{CR}(\boldsymbol{y}_a^{(J)}, \boldsymbol{y}_p^{(J)}), \tag{6}$$

where $\boldsymbol{y}_a$ and $\boldsymbol{y}_p$ are the auxiliary and primary prediction as previous work [22, 34] do, and $\mathcal{L}_{CR}$ is the complementary regularization [27] that proposed to handle co-adaptation noises.

As mentioned in Sect. 3.2, we have projected both video and skeleton data into the same graph-structured space ($\boldsymbol{V}^{(0)}$ and $\boldsymbol{J}^{(0)}$) and can use the same architecture for both types of inputs. Therefore, the architecture and supervision of the slow pathway are the same as the fast pathway ($\mathcal{L}_{slow} = \mathcal{L}_{fast} = \mathcal{L}_{basic}$). To effectively utilize the complementary information in both pathways, we propose group-wise cross-attention and frame-wise fusion to progressively fuse the intermediate features.

Group-wise Cross-attention. The inaccurate estimation of joints often occurs with hands due to the inherent flexibility of hand movements, which necessitates an early integration of skeleton and video data to rectify such inaccuracies. To address this, we design a group-wise cross-attention operation to fuse features from both pathways subsequent to each group-specific GCN layer. Specifically, the fast-to-slow fusion process can be formulated as:

$$V_{t,j}^{(i)} = V_{t,j}^{(i)} + \sum_{k \in G(j)} \psi(V_{t,j}^{(i)}, J_{\pi_t,k}^{(i)}) \phi(J_{\pi_t,k}^{(i)}), \tag{7}$$

where $G(j)$ denotes the joint set belonging to the same group as joint j, and π_t represents the index of the temporally closest frame in the fast pathway. $\psi(\cdot,\cdot)$ computes the similarity between two feature vectors and ϕ denotes a projection function. Analogously, we employ a slow-to-fast fusion in the reverse direction.

Frame-wise Fusion. As shown in Fig. 1c, we adopt independent contextual modules for slow and fast pathways. To better take advantage of both pathways, we utilize an extra contextual module that adopts the frame-wise fused features as input and is supervised by:

$$\mathcal{L}_{extra} = \mathcal{L}_{CTC}(y_a^{(F)}) + \mathcal{L}_{CTC}(y_p^{(F)}), \tag{8}$$

where $y_a^{(F)}$ and $y_p^{(F)}$ are auxiliary and primary predictions of the fusion pathway.

3.4 Multi-Projector Distillation

We have projected both video and skeleton data into the same graph-structured space and progressively fuse them with S^{2}Net. We further explore the possibility

of enhancing inference efficiency by distilling knowledge learned from an ensemble of pathways into a skeleton-only model. Considering potential inconsistencies in label sequence alignment across pathways, we adopt a multi-projector distillation approach similar to prior work [8]. This method involves attaching three additional projectors to the convolutional features of the skeleton-only model, supervising the distillation process with:

$$\mathcal{L}_{distill} = \mathcal{L}_{basic} + \sum_{\Gamma \in \{J,V,F\}} \lambda_{KL} KL(\tilde{y}_{proj}^{(\Gamma)}, y_p^{(\Gamma)}), \tag{9}$$

where $\tilde{y}_{proj}^{(\Gamma)}$ is the corresponding projector for the Γ pathway.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate S^{2}Net across multiple datasets: Phoenix14, Phoenix14T, and CSL-Daily. We utilize the Word Error Rate (WER) as the evaluation metric, where a lower WER indicates better recognition performance.

- ◇ **Phoenix14** [29] is a video dataset, which comprises a total of 1295 sign language phrases, organized into 6841 sentences, including 5672 training samples, 540 development samples (Dev set), and 629 testing samples (Test set).
- ◇ **Phoenix14T** [5] is an extended version of the Phoenix14. This dataset includes 1085 sign language phrases, totaling 8247 sentences, including 7096 training samples, 519 development samples, and 642 testing samples.
- ◇ **CSL-Daily** [55] is a comprehensive dataset, which comprises 2000 sign language phrases, amounting to 20654 sentences. These are divided into 18401 training samples, 1077 development samples, and 1176 testing samples.

Baseline. The slow pathway employs an index-based projector to generate a graph-structured representation. Both the fast and slow pathways are trained independently. Each pathway employs a stacked three-layer group-wise ST-GCN [27] as the visual extractor. Similar to previous methods [27,34], the contextual module adopts a 1D-CNN (K3-P2-K3-P2) along with a two-layer BiLSTM, where Kδ and Pδ denote the 1D convolutional and the max pooling layers with a kernel size of δ, respectively. The hidden layer dimension of the two-layer BiLSTM is set to 1024. Finally, the outputs from both pathways are averaged as a straightforward fusion approach.

Implementation Details. For the slow pathway, graph-structured representation is derived from mid-level feature maps extracted after the first two layers of a ResNet18 model pre-trained on the relevant sign language dataset based on SMKD [22]. For the fast pathway, we use MMPose to obtain the whole body skeleton data as previous work does [27]. The entire model is trained for 40 epochs with a minimum batch size of 4 and optimized by the AdamW optimizer with a weight decay of 1×10^{-4}. More details can be found in the supplementary material.

Table 1: Performance comparison (WER %) on Phoenix14/14-T and CSL-Daily. The highest performance within each group is highlighted in **bold**.

Method	PHOENIX14		PHOENIX14-T		CSL-Daily	
	Dev	Test	Dev	Test	Dev	Test
RGB-based						
DNF [13]	23.8	24.4	-	-	32.8	32.4
VAC [34]	21.2	22.3	-	-	33.3	32.6
CMA [37]	21.3	21.9	-	-	-	-
SignBT [55]	-	-	22.7	23.9	33.2	33.2
SMKD [22]	20.8	21.0	20.8	22.4	28.4	27.5
TLP [25]	19.7	20.8	19.4	21.2	-	-
CTCA [21]	19.5	20.1	19.3	20.3	31.3	29.4
CorrNet [26]	18.8	19.4	18.9	20.5	30.6	30.1
SlowFastSign [1]	18.0	**18.3**	17.7	**18.7**	25.5	**24.9**
Skeleton-based						
SignBERT+ [24]	34.0	34.1	32.9	33.6	-	-
TwoStream-SLR [9]	28.6	28.0	27.1	27.2	34.6	34.1
CoSign [27]	19.7	**20.1**	19.5	**20.1**	28.1	**27.2**
Fusion-based (RGB+language)						
CVT-SLR [53]	19.8	20.1	19.4	20.3	-	-
C^2ST [50]	17.5	**17.7**	17.3	**18.9**	25.9	**25.8**
Fusion-based (RGB+skeleton)						
STMC [56]	21.1	20.7	19.6	21.0	-	-
C^2SLR [57]	20.5	20.4	20.2	20.4	-	-
TwoStream-SLR [9]	18.4	18.8	17.7	19.3	25.4	25.3
Baseline	19.3	19.4	18.6	19.5	27.7	27.2
S^2Net(Index-based)	17.6	**17.5**	17.7	**18.4**	25.4	**24.5**
S^2Net(Query-based)	17.4	**17.5**	17.2	18.9	25.8	**24.5**

4.2 Comparison with State-of-the-art Methods

We evaluate the proposed model against previous CSLR methods on Phoenix14, Phoenix14-T, and CSL-Daily datasets. As shown in Table 1, the proposed S^2Net achieves competitive results across all datasets. The baseline outperforms the previous skeleton-based SOTA method CoSign [27] with a 0.4%/0.7% reduction in WER on Phoenix14, highlighting the importance of complementary video data for sign language recognition. Additionally, we evaluate two types of projectors for generating graph-structured representations, both of which significantly improve recognition. The Index-based projector reduces WER than baseline by 1.7%/1.9%, 0.9%/1.1%, and 2.3%/2.7% on the Phoenix14, Phoenix14-T, and CSL-Daily datasets, demonstrating the effectiveness of the proposed method.

When compared with TwoStream-SLR [9] under similar input conditions, S^2Net achieves a WER reduction of 1.0%/1.3% on Phoenix14, with similar trends observed in the other two datasets. Furthermore, S^2Net shows a 0.6%/0.8% improvement in recognition on Phoenix14 over SlowFastSign [1], which relies

Table 2: Efficiency comparison excluding the pose estimation stage on Phoenix14. The computational cost is measured in frame-wise FLOPs (FLoating-point OPerations) across different architectures.

Method	Backbone	Frame-wise FLOPs(G)	PHOENIX14	
			Dev	Test
TwoStream-SLR [9]	S3D (modified)	2.2	18.4	18.8
SlowFastSign [1]	SlowFast101	3.7	18.0	18.3
S^2Net(Index-based)	Resnet18+GCN	**0.6**	17.6	**17.5**
S^2Net(Query-based)	Resnet18+GCN	0.7	17.4	**17.5**

Table 3: Ablation results (WER %) of index layer on Phoenix14. '-' denotes using the entire ResNet18 as the visual extractor without indexing. The index layer number indicates the number of residual layers from ResNet18 utilized.

Freeze Weight	Index Layer	ImageNet Pre-trained		Phoenix14 Pre-trained	
		Dev	Test	Dev	Test
	-	**20.7**	21.7	20.3	20.7
✓	-	21.5	21.7	**20.2**	20.8
✓	1	**20.8**	20.8	20.6	20.9
✓	2	21.9	22.7	**19.3**	19.4
✓	3	25.6	26.0	24.2	24.1
✓	4	29.5	29.7	29.0	29.7

solely on video input. This improvement is attributed to the effective extraction of projecting both video and skeleton data into the same graph-structured space.

Moreover, we provide an efficiency comparison between the proposed method, TwoStream-SLR [9] and SlowFastSign [1] in Table 2. TwoStream-SLR utilizes only the first four blocks of S3D, resulting in approximately 2.2 GFLOPs of the summation frame-wise FLOPs of two streams. SlowFastSign adopts a much heavier backbone, which is a modified version of ResNet101, and the frame-wise FLOPs is $3.7 \approx 234/64$ G (approximately 234 GFLOPs for each 64-frame clip) according to SlowFast [17]. The proposed method achieves significant performance gain ($\geq$ 0.8 WER on the test set) while maintaining higher efficiency, with frame-wise FLOPs of approximately 0.6G (100.7 GFLOPs for an average length of 159 frames).

4.3 Ablation Studies

Ablation on the abstract level of visual features. As mentioned in Sect. 3.2, the slow pathway incorporates a pre-trained visual module to construct the graph-structured representation. To enhance the efficiency of the training process, we first evaluate the effects of the pre-trained dataset and the abstract level of visual features, and present results in Table 3. Using a visual extractor pre-trained on Phoenix14 generally yields better performance across most set-

Table 4: Ablation results (WER %) of S^{2}Net on Phoenix14, GCA and FF denote Group-wise Cross-attention and Frame-wise Fusion, respectively.

Graph-structured Projector		GCA	FF	Dev	Test
Index-based	Query-based				
✓				19.3	19.4
✓		✓		17.8	18.2
✓			✓	18.1	18.5
✓		✓	✓	**17.7**	17.7
	✓	✓	✓	**17.4**	17.5

Table 5: Ablation results (WER %) of temporal stride on Phoenix14.

Temporal Stride	Fast	Fusion	FLOPs	Training Memory	Training Time
$\boldsymbol{\alpha} = 1$	**18.2**/18.0	**17.5**/17.7	184.8 G	15.0 GB	∼4.5h
$\boldsymbol{\alpha} = 2$	18.5/18.7	17.7/17.7	100.7 G	11.0 GB	∼4.0h
$\boldsymbol{\alpha} = 4$	19.4/19.0	19.0/18.8	57.7 G	10.4 GB	∼3.7h
$\boldsymbol{\alpha} = 8$	19.5/19.9	21.1/21.4	36.5 G	9.7 GB	∼3.5h

tings compared to using one pre-trained on ImageNet. Additionally, whether the visual extractor is frozen or not has minimal impact on recognition performance.

Furthermore, we observe that indexing the extracted features after the second residual block of a ResNet18 pre-trained on Phoenix14 yields better performance than using the entire model (reducing WER from 20.2% to 19.3%). Considering training efficiency, we pre-extract features from sign language videos using the visual extractor pre-trained on the Phoenix14 dataset and use these features as input for the remaining experiments without further specification.

Ablation on S^{2}Net designs. To demonstrate the effectiveness of the proposed S^{2}Net, we conduct ablation on the design of S^{2}Net and present the results in Table 4. Compared to the straightforward fusion approach, adopting either group-wise cross-attention or frame-wise fusion can improve the recognition results, and adopting both can bring further improvements (from 19.3% to 17.7% on the dev set). Moreover, adopting the query-based projector achieves competitive performance compared to the index-based projector, which indicates the learnable embeddings effectively capture corresponding visual features similar to those indexed by the pose.

Ablation on temporal stride. In the slow pathway, we conduct ablation on the temporal stride of α to balance efficiency and accuracy. As shown in Table 5, the performance continuously improves with the gradual decrease of the temporal stride, while the computational overhead and the memory usage during training increase correspondingly. Notably, when the temporal stride is reduced to 2, the WER on the Dev set remains nearly identical to that at a stride of 1, differing by only 0.2%, but the training cost is significantly reduced (from 184.8 GFLOPs to

Table 6: Ablation results (WER %) of multi-projector distillation on Phoenix14. nT denotes the best performance using n pathways for distillation.

Projector	Student Performance				Teacher Performance
	0T	1T	2T	3T	
✓	20.4/21.0	**19.7**/20.4 19.9/20.5	19.8/20.1 20.0/20.0	19.8/19.9 20.0/20.2	17.7/17.7

Table 7: Ablation results (WER %) of robustness to Gaussian noises on Phoenix14.

Projector	Gaussian Augmentation	Gaussian Noise				
		0.0	1.5	3.0	5.0	8.0
Index-based		17.6	86.8	98.7	99.7	100.0
	✓	**17.4**	17.9	22.6	33.2	67.6
Query-based		**17.4**	99.1	98.9	98.7	98.6
	✓	**17.4**	18.2	23.0	32.0	65.1

100.7 GFLOPs with the average length of 159 frames). Considering the balance between efficiency and accuracy, we adopt $\alpha = 2$ as the default setting of S^2Net.

Ablation on multi-projector distillation. We further adopt the trained S^2Net as the teacher model and explore the effects of multi-projector distillation to a skeleton-only student model. As presented in Table 6, distillation with a single pathway achieves the best performance, which is competitive with CoSign-2s [27] (19.7%/20.4% vs. 19.7%/20.1%) with less computation cost (5.8 GFLOPs vs. 30.1 GFLOPs). Adopting an extra projector for distillation can consistently reduce the WER (about 0.2% on the dev set).

Ablation on robustness to noise. We evaluate the robustness of different projectors for the Gaussian noises and present the results in Table 7. The noise is categorized into five levels, with levels three (3.0) and five (8.0) causing disturbances where the hand and body are difficult to distinguish. Both types of projectors suffer significant performance losses and achieve competitive performance when adopting the Gaussian augmentation during training. The index-based projector performs better under minor noise, while the query-based projector outperforms as noise intensity increases. Besides, we notice that the slow pathway of the query-based method is more robust than that of the index-based (81.8/87.6 vs. 85.1/89.4 for Gaussian noises 3.0/8.0).

4.4 Qualitative Experiments

To provide a more intuitive understanding of the proposed methods, we visualize the accumulated cross-attention map of the query-based projector and the index map of the index-based projector in Fig. 3. The query-based projector can localize the hand region more accurately, especially during fast hand movements or instances of occlusion. Moreover, the query-based projector comprehensively

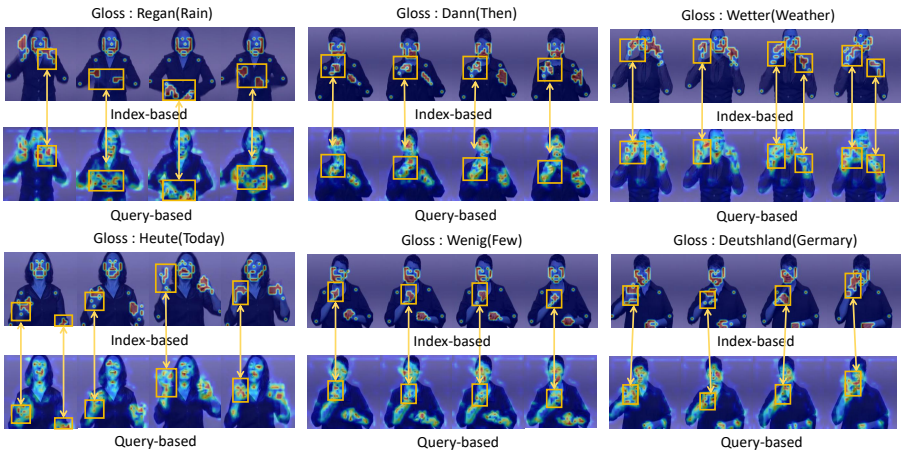


Fig. 3: Visualization of cross-attention maps for query-based projector and the index maps for index-based projector. The highlighted areas within the boxes indicate regions where the index-based projector is inaccurate. **Best view in color**

captures fine-grained details, with particular emphasis on the signer's arms and neck, which makes it effective for analyzing complex gestures.

5 Conclusion

In this paper, we propose a Skeleton-aware SlowFast Network (S^2Net) for continuous sign language recognition. S^2Net projects video and skeleton data into a unified graph-structured space through either index-based or query-based projectors, and employs a consistent GCN-based architecture for both pathways. The proposed method adopts a group-wise cross-attention module and frame-wise fusion to progressively fuse complementary information from different pathways. Experimental results demonstrate that projecting sign videos into a graph-structured representation effectively enhances the complementary information available from skeleton data. By progressively integrating both pathways, competitive performance is achieved, which can be further distilled into a skeleton-only model. We hope the proposed method can inspire further research into efficient and effective representation learning in CSLR.

Acknowledgments

This work is partially supported by the National Science and Technology Major Project (No. 2021ZD0111901) and the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20240762.

References

1. Ahn, J., Jang, Y., Chung, J.S.: Slowfast network for continuous sign language recognition. In: ICASSP. pp. 3920–3924. IEEE (2024) 4, 5, 10, 11
2. Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J.S., Fox, N., Zisserman, A.: Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In: Eur. Conf. Comput. Vis. pp. 35–53. Springer (2020) 3, 4
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Int. Conf. Comput. Vis. pp. 6836–6846 (2021) 4
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Eur. Conf. Comput. Vis. pp. 561–578. Springer (2016) 1, 4, 5
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10023–10033 (2020) 9
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7291–7299 (2017) 1, 4, 5
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6299–6308 (2017) 3, 4
8. Chen, Y., Wang, S., Liu, J., Xu, X., de Hoog, F., Huang, Z.: Improved feature distillation via projector ensemble. Adv. Neural Inform. Process. Syst. **35**, 12084–12095 (2022) 9
9. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. Advances in Neural Information Processing Systems **35**, 17043–17056 (2022) 2, 3, 4, 6, 10, 11
10. Cheng, K.L., Yang, Z., Chen, Q., Tai, Y.W.: Fully convolutional networks for continuous sign language recognition. In: Eur. Conf. Comput. Vis. pp. 697–714. Springer (2020) 3
11. Cihan Camgoz, N., Hadfield, S., Koller, O., Bowden, R.: Subunets: End-to-end hand shape and continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 3056–3065 (2017) 3
12. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020) 6
13. Cui, R., Liu, H., Zhang, C.: A deep neural framework for continuous sign language recognition by iterative training. IEEE Trans. Multimedia **21**(7), 1880–1891 (2019) 3, 10
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021) 3
15. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2969–2978 (2022) 2, 4, 6
16. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 203–213 (2020) 5
17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Int. Conf. Comput. Vis. pp. 6202–6211 (2019) 4, 7, 11

18. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. Adv. Neural Inform. Process. Syst. **35**, 35946–35958 (2022) 4, 5
19. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Int. Conf. Mach. Learn. pp. 369–376 (2006) 4
20. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7297–7306 (2018) 1, 4, 5
21. Guo, L., Xue, W., Guo, Q., Liu, B., Zhang, K., Yuan, T., Chen, S.: Distilling cross-temporal contexts for continuous sign language recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10771–10780 (2023) 3, 10
22. Hao, A., Min, Y., Chen, X.: Self-mutual distillation learning for continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 11303–11312 (2021) 3, 8, 9, 10
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016) 3
24. Hu, H., Zhao, W., Zhou, W., Li, H.: Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 11221–11239 (2023) 3, 10
25. Hu, L., Gao, L., Liu, Z., Feng, W.: Temporal lift pooling for continuous sign language recognition. In: Eur. Conf. Comput. Vis. pp. 511–527. Springer (2022) 10
26. Hu, L., Gao, L., Liu, Z., Feng, W.: Continuous sign language recognition with correlation network. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2529–2539 (2023) 3, 10
27. Jiao, P., Min, Y., Li, Y., Wang, X., Lei, L., Chen, X.: Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 20676–20686 (2023) 2, 3, 5, 6, 7, 8, 9, 10, 13
28. Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language. arXiv preprint arXiv:1812.01053 (2018) 3
29. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 9
30. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3793–3802 (2016) 3
31. Koller, O., Zargaran, O., Ney, H., Bowden, R.: Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In: Brit. Mach. Vis. Conf. (2016) 3
32. Koller, O., Zargaran, S., Ney, H.: Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4297–4305 (2017) 3
33. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: Winter Conf. on App. of Comput. Vis. pp. 1459–1469 (2020) 3
34. Min, Y., Hao, A., Chai, X., Chen, X.: Visual alignment constraint for continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 11542–11551 (2021) 3, 5, 8, 9, 10
35. Niu, Z., Mak, B.: Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In: Eur. Conf. Comput. Vis. pp. 172–186. Springer (2020) 3

36. Parelli, M., Papadimitriou, K., Potamianos, G., Pavlakos, G., Maragos, P.: Spatio-temporal graph convolutional networks for continuous sign language recognition. In: ICASSP. pp. 8457–8461. IEEE (2022) 2, 3
37. Pu, J., Zhou, W., Hu, H., Li, H.: Boosting continuous sign language recognition via cross modality augmentation. In: ACM Int. Conf. Multimedia. pp. 1497–1505 (2020) 10
38. Pu, J., Zhou, W., Li, H.: Dilated convolutional network with iterative optimization for continuous sign language recognition. In: IJCAI. vol. 3, p. 7 (2018) 3
39. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: Int. Conf. Comput. Vis. pp. 5533–5541 (2017) 4
40. Schiappa, M.C., Rawat, Y.S., Shah, M.: Self-supervised learning for videos: A survey. ACM Computing Surveys **55**(13s), 1–37 (2023) 4
41. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1010–1019 (2016) 4
42. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Adv. Neural Inform. Process. Syst. **27** (2014) 4, 5
43. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: A review. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3200–3225 (2022) 2, 4
44. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7794–7803 (2018) 4
45. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2566–2576 (2019) 4
46. Wei, D., Lim, J.J., Zisserman, A., Freeman, W.T.: Learning and using the arrow of time. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8052–8060 (2018) 4
47. Wei, F., Chen, Y.: Improving continuous sign language recognition with cross-lingual signs. In: Int. Conf. Comput. Vis. pp. 23612–23621 (2023) 3, 4
48. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Eur. Conf. Comput. Vis. pp. 305–321 (2018) 3, 4
49. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI. vol. 32 (2018) 4, 7
50. Zhang, H., Guo, Z., Yang, Y., Liu, X., Hu, D.: C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 21053–21062 (2023) 3, 10
51. Zhao, W., Hu, H., Zhou, W., Mao, Y., Wang, M., Li, H.: Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology (2024) 3
52. Zhao, W., Zhou, W., Hu, H., Wang, M., Li, H.: Self-supervised representation learning with spatial-temporal consistency for sign language recognition. arXiv preprint arXiv:2406.10501 (2024) 3
53. Zheng, J., Wang, Y., Tan, C., Li, S., Wang, G., Xia, J., Chen, Y., Li, S.Z.: Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23141–23150 (2023) 3, 4, 10
54. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Int. Conf. Comput. Vis. pp. 20871–20881 (2023) 4

55. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1316–1325 (2021) 9, 10
56. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. In: AAAI. vol. 34, pp. 13009–13016 (2020) 2, 3, 4, 10
57. Zuo, R., Mak, B.: C2slr: Consistency-enhanced continuous sign language recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5131–5140 (2022) 2, 3, 4, 10
58. Zuo, R., Wei, F., Mak, B.: Natural language-assisted sign language recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14890–14900 (2023) 3, 4